

引言

(1) 随机现象可以用什么样的分布律 (分布函数) 来刻画, 这种分布的选择合理吗?

(2) 所选用的分布的参数是多少? 如何估计和确定这些参数?

要解决上述问题, 只能求助于观测, 合理地取得一些数据, 据此作出统计上的推断, 从而解决问题. 这就是数理统计的基本且主要任务.

更准确地说 数理统计的主要内容是：

1. **实验设计和研究**，即研究如何更合理、更有效地抽取样本来获得观测数据和资料的方法；

2. **统计推断**：如何利用一定的数据资料，对所关心的问题，得出尽可能准确的统计结论：

(1) **估计**——从局部观测资料的统计特征，推断所观测对象的总体特征（包括总体分布与数字特征）；

(2) **假设检验**——依据抽样数据资料，对总体的某种假设做检验，从而决定对此假设是拒绝还是接受。

例 某钢筋厂日产某型号钢筋10000根，质量检验员每天只抽查50根的强度，于是提出以下问题：

(1) 如何从仅有的50根钢筋的强度数据去估计整批（10000根）钢筋的强度平均值？又如何估计这批钢筋强度偏离平均值的离散程度？

----参数估计

(2) 若规定了这种型号钢筋的标准强度，从抽查得的50个强度数据如何判断整批钢筋的平均强度与规定标准有无差异？

----假设检验

第6章 数理统计的基本概念

6.1 总体与样本

6.2 样本数据的整理

6.3 统计量

6.4 抽样分布

6.1 总体与样本

1. 总体

在数理统计中，研究对象的某个数量指标的全体组成的集合. 该数量指标是一个随机变量（或者多维随机变量），记作 X .

2. 个体

组成总体的每一个元素，即总体的每个数理指标，可看作随机变量 X 的某个取值，用 X_i 表示.

3. 样本

从总体中抽取的部分个体组成的向量，用 $(X_1, \dots, X_n)$ 表示， n 称为样本容量. 称 $(x_1, \dots, x_n)$ 为总体 X 的容量为 n 的样本观测值. 样本具有二重性，即**随机变量**和**具体的观测值**.

4. 简单随机样本

若总体 X 的样本 $(X_1, \dots, X_n)$ 满足:

- (1) 样本的每个分量 $X_i (i = 1, 2, \dots, n)$ 必须与总体 X 具有相同的分布;
- (2) 样本的各个分量 $X_1, \dots, X_n$ 是相互独立的.

凡具有上述两个性质的样本称为简单随机样本. 获得简单随机样本的抽样方法称为简单随机抽样.

对于简单随机样本，其概率分布可以由总体 X 的分布完全决定，设总体 X 的分布函数为 $F(x)$ ，则样本 $(X_1, \dots, X_n)$ 的分布函数为

$$F(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F(x_i)$$

若总体 X 的概率密度函数为 $f(x)$ ，则样本 $(X_1, \dots, X_n)$ 的联合概率密度函数为

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i)$$

例 1 设 X 服从参数为 λ ($\lambda > 0$) 的指数分布, $(X_1, \dots, X_n)$ 是来自总体 X 的样本, 求样本 $(X_1, \dots, X_n)$ 的概率密度

解: 因为 X 服从指数分布, 则总体 X 的概率密度为

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

因为 $X_1, \dots, X_n$ 相互独立, 且都与 X 有相同的分布, 所以 $(X_1, \dots, X_n)$ 的概率密度为

$$\begin{aligned} f_n(x_1, x_2, \dots, x_n) &= \prod_{i=1}^n f(x_i) \\ &= \begin{cases} \lambda^n e^{-\lambda \sum_{i=1}^n x_i}, & x_1, x_2, \dots, x_n \geq 0 \\ 0, & \text{其他} \end{cases} \end{aligned}$$

如果总体 X 是离散型随机变量，具有分布律

$$P\{X = a_i\} = p_i \quad (i = 1, 2, \dots)$$

则样本 $(X_1, \dots, X_n)$ 的分布律为

$$\begin{aligned} P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\} &= \prod_{i=1}^n P\{X_i = x_i\} \\ &= \prod_{i=1}^n P\{X = x_i\} \end{aligned}$$

其中 $x_1, x_2, \dots, x_n$ 都在集合 $\{a_1, a_2, \dots\}$ 中取值

例 2 设总体 X 服从 $(0-1)$ 分布 $B(1, p)$, 其中 $0 < p < 1$,
($X_1, \dots, X_n$) 是来自总体 X 的样本, 求样本 ($X_1, \dots, X_n$) 的
分布律

解: 因为 X 服从指数分布, 则总体 X 的分布律为

$$P\{X = i\} = p^i (1-p)^{1-i}, \quad i = 0, 1$$

因为 $X_1, \dots, X_n$ 相互独立, 且都与 X 有相同的分布, 所以
($X_1, \dots, X_n$) 的分布律为

$$\begin{aligned} P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\} &= \prod_{i=1}^n P\{X_i = x_i\} \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i} \end{aligned}$$

其中 $x_1, x_2, \dots, x_n$ 都在集合 $\{0, 1\}$ 中取值

6.2 样本数据的整理

1. 样本频数分布与频率分布

设有样本值 $(x_1, x_2, \dots, x_n)$ 样本频数分布是指样本值中不同数值在样本值中出现的**频次（即次数）**，样本频率分布是指样本值中不同数值在样本值中出现的**频率（即频数除以样本容量）**

- 想象一个随机试验，例如投掷两枚硬币并记录正面出现的次数，用 X 表示，则 X 就是随机变量，将两枚硬币试验重复1000次，得到

$p(x)$	x	频数(m_x)	频率(m_x/n)
0.25	0	260	0.260
0.5	1	517	0.517
0.25	2	223	0.223

- 对于离散型随机变量，当 n 很大时，**样本频率分布作为总体分布律**
- 如果是连续型呢？ **频率直方图**

2. 频率直方图

对连续总体 X ,样本分布通常用频率直方图来描述,并由此近似得出总体 X 的概率密度函数 $f(x)$ 的曲线.

画频率直方图步骤:

- 1) 整理数据;
- 2) 分组;
- 3) 列分组频率分布表;
- 4) 作频率直方图;
- 5) 作概率密度曲线

➤ 某轧钢厂生产了一批钢材，为了研究这批钢材的抗拉强度，从中随机抽取了76个样本进行试验

抗拉强度观测值											
41.0	37.0	33.0	44.2	30.5	27.0	45.0	28.5	31.2	33.5	38.5	41.5
43.0	45.5	42.5	39.0	38.8	35.5	32.5	29.6	32.6	34.5	37.5	39.5
42.8	45.1	42.8	45.8	39.8	37.2	33.8	31.2	29.0	35.2	37.8	41.2
43.8	48.0	43.6	41.8	36.6	34.8	31.0	32.0	33.5	37.4	40.8	44.7
40.2	41.3	38.8	34.1	31.8	34.6	38.3	41.3	30.0	35.2	37.5	40.5
38.1	37.3	37.1	41.5	29.5	29.1	27.5	34.8	36.5	44.2	40.0	44.5
40.6	36.2	35.8	31.5								

1) **整理数据** 将样本值 $x_1, x_2, \dots, x_n$ 按从小到大的顺序排列

$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 这样排不仅可以看出其最大值与最小值, 还可以看出样本值的范围.

2) **数据分组** 在包含 $[x_{(1)}, x_{(n)}]$ 的区间 $[a, b]$ 中插入一些分点, 把 $[a, b]$ 分成一些小区间; 每个小区间的长度称为**组距**; 区间的中点称为**组中值**; 小区间的个数 l 称为**组数**. 一般采用等分, 即各组的组距相等. 组数 l 的大小根据样本容量而定.

3) **列分组频率分布表** 令 m_i 表示观察值落入 $(t_i, t_{i+1}]$ 中的个数, 则 $f_i = m_i/n$ 为该组的频率, 由大数定律知频率 f_i 近似等于总体落入区间 $(t_i, t_{i+1}]$ 的概率.

$$f_i \approx p_i = \int_{t_{i-1}}^{t_i} f(x) dx = f(\xi_i) d_i$$

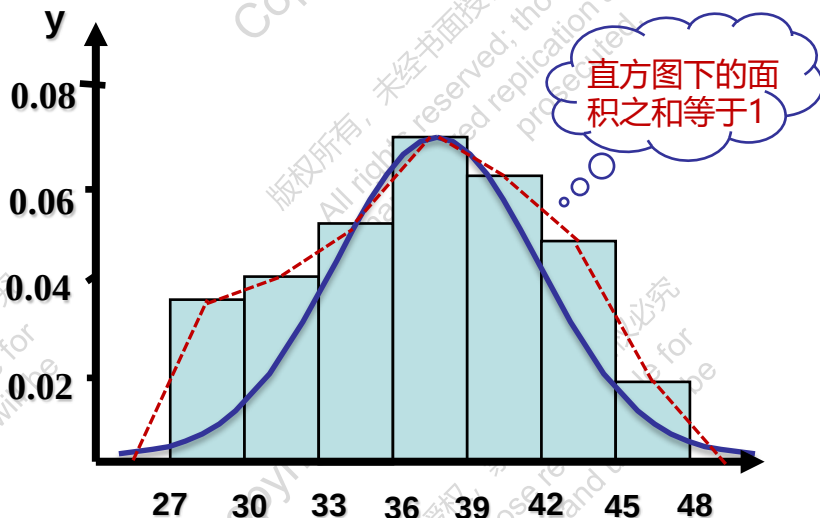
$$y_i = f_i / d_i \approx f(\xi_i)$$

分组	组中值	频数 m_i	频率 f_i	$y_i = f_i/d_i$
[27,30]	28.5	8	0.105	0.035
(30,33]	31.5	10	0.132	0.044
(33,36]	34.5	12	0.158	0.053
(36,39]	37.5	17	0.224	0.074
(39,42]	40.5	14	0.184	0.061
(42,45]	43.5	11	0.145	0.048
(45,48]	46.5	4	0.053	0.018

4) 作出频率直方图

在 xoy 平面上,分别以各区间 $(t_{i-1}, t_i]$ 为底,以 $y_i = f_i/d_i$ 为高画一排竖着的矩形,即为频率直方图;

5) 作概率密度曲线



把频率直方图中每个矩形边上的中点连接起来,得到一条折线,当 n 及 l 充分大时,可近似认为是 x 的概率密度函数曲线.

3. 经验分布函数

样本频率分布和样本频率直方图分别只适用于总体 X 为离散型和连续型的随机变量情形. 经验分布函数法是一种更为普遍的估计总体分布的方法.

设有样本 $(x_1, x_2, \dots, x_n)$. 作函数

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{\{x_i \leq x\}}(x)$$

其中 $I_A(\cdot)$ 表示集合 A 的示性函数, $I_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$ 显然, 函数 $F_n(x)$ 具有如下性质

- (1) $F_n(x)$ 是 x 的单调增函数
- (2) $F_n(x)$ 是 x 的右连续函数
- (3) $F_n(-\infty) = 0, F_n(+\infty) = 1$

因此, $F_n(x)$ 是一个分布函数, 即经验分布函数

如果样本值，是以频数分布表给出，那么 $F_n(x)$ 可具体表达为

$$F_n(x) = \begin{cases} 0, & x < x_1^*, \\ \frac{m_1 + m_2 + \cdots + m_i}{n}, & x_i^* \leq x < x_{i+1}^*, (i = 1, 2, \cdots, l) \\ 1, & x \geq x_l^*; \end{cases}$$

由大数定律，对于每个固定 x 值，当 $n \rightarrow \infty$ 时，有

$$\lim_{n \rightarrow \infty} P\{|F_n(x) - F(x)| < \varepsilon\} = 1 \quad (\forall \varepsilon > 0)$$

例 3 从总体 X 中抽取容量为7的样本，其观测值为：32, 65, 28, 32, 35, 30, 29. 试求 X 的经验分布函数.

解： 将样本值由小到大排序得 $28 < 29 < 30 < 32 = 32 < 35 < 65$

$$F_7(x) = \begin{cases} 0, & x < 28, \\ \frac{1}{7}, & 28 \leq x < 29, \\ \frac{2}{7}, & 29 \leq x < 30, \\ \frac{3}{7}, & 30 \leq x < 32, \\ \frac{5}{7}, & 32 \leq x < 35, \\ \frac{6}{7}, & 35 \leq x < 65, \\ 1, & 65 \leq x; \end{cases}$$

6.3 统计量

1. 统计量的概念

- 由样本推断总体特征，需要对样本进行“加工”、“提炼”。这就需要构造一些样本的函数，它把样本中所含的信息集中起来。

定义6.1:

设 $(X_1, X_2, \dots, X_n)$ 是来自总体 X 的一个样本, $T = g(X_1, X_2, \dots, X_n)$ 为 $(X_1, X_2, \dots, X_n)$ 的一个实值函数, 且 g 中不包含任何未知参数, 则称 T 为样本 $(X_1, X_2, \dots, X_n)$ 的一个统计量. 若 $(x_1, x_2, \dots, x_n)$ 是样本 $(X_1, X_2, \dots, X_n)$ 的一个观测值, 则 $t = g(x_1, x_2, \dots, x_n)$ 称为统计量 T 的一个观测值.

2. 几个常用的统计量

设 $(X_1, X_2, \dots, X_n)$ 是来自总体 X 的样本, $(x_1, x_2, \dots, x_n)$ 是这一样本的观测值.

$$(1) \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} (X_1 + X_2 + \dots + X_n) \quad \text{---样本均值}$$

$$(2) \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{---样本方差}$$

$$(3) \quad S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \quad \text{---样本标准差}$$

$$(4) \quad A_k = \frac{1}{n} \sum_{i=1}^n X_i^k \quad \text{---样本的} k \text{阶原点矩}$$

$$(5) \quad B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k \quad \text{---样本的} k \text{阶中心矩}$$

写出对应的观测值?

定理6.1 若总体 X 的均值 (数学期望) $E(X) = \mu$ 和方差 $D(X) = \sigma^2$ 存在, 则

$$(1) E(\bar{X}) = \mu, \quad D(\bar{X}) = \frac{\sigma^2}{n}$$

$$(2) (p)\lim_{n \rightarrow \infty} \bar{X} = \mu$$

证: (1) 因 $X_1, X_2, \dots, X_n$ 都与 X 有相同的同分布, 所以

$E(X_i) = \mu, \quad D(X_i) = \frac{\sigma^2}{n} (i = 1, 2, \dots, n).$ 又因 $X_1, X_2, \dots, X_n$ 相互独立, 所以

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu,$$

$$D(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n D(X_i) = \frac{\sigma^2}{n}$$

(2) 由辛钦大数定律知成立.

定理6.2 若总体 X 的均值 $E(X) = \mu$ 和方差 $D(X) = \sigma^2$ 存在, 则

$$(1) E(S^2) = \sigma^2$$

$$(2) (p) \lim_{n \rightarrow \infty} S^2 = \sigma^2$$

证: (1) 由定理6.1知, $E(\bar{X}) = \mu$, $D(\bar{X}) = \frac{\sigma^2}{n}$,

$$E(\bar{X}^2) = D(\bar{X}) + E(\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2$$

因 $X_1, X_2, \dots, X_n$ 都与 X 有相同的同分布, 且相互独立, 所以, $X_1^2, X_2^2, \dots, X_n^2$ 相互独立, 且都与 X^2 同分布, 从而

$$E(X_i^2) = D(X_i) + E(X_i)^2 = \sigma^2 + \mu^2$$

$$\begin{aligned}
 \text{故 } E(S^2) &= E\left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right) \\
 &= \frac{1}{n-1} \left[\sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2) \right] \\
 &= \frac{1}{n-1} \left\{ \sum_{i=1}^n (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \right\} \\
 &= \frac{1}{n-1} [n(\sigma^2 + \mu^2) - \sigma^2 - n\mu^2] = \sigma^2.
 \end{aligned}$$

(2) 对 $X_1^2, X_2^2, \dots, X_n^2$ 应用辛钦大数定律, 有

$$(p) \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i^2 = E(X^2) = \sigma^2 + \mu^2$$

由定理6.1知

$$(p) \lim_{n \rightarrow \infty} \bar{X} = \mu$$

故由5.1节中依概率收敛序列的性质，得

$$\begin{aligned}(p)\lim_{n \rightarrow \infty} S^2 &= (p)\lim_{n \rightarrow \infty} \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 \right) \\ &= \sigma^2 + \mu^2 - \mu^2 = \sigma^2\end{aligned}$$

定理6.3 (1) 若总体 X 的 k 阶原点矩 $\alpha_k = E(X^k)$ 存在，其中 k 为正整数，则

$$E(A_k) = \alpha_k$$

$$(2) (p)\lim_{n \rightarrow \infty} A_k = \alpha_k.$$

证明与上面定理证明类似，故省略。

3. 顺序统计量

- 设 $(X_1, X_2, \dots, X_n)$ 是来自总体 X 的样本, $(x_1, x_2, \dots, x_n)$ 是这一样本的一个观测值, 将观测值 $x_1, x_2, \dots, x_n$ 由小到大排序为

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

定义统计量 $X_{(k)}$ 取值为 $x_{(k)}$, $k=1, 2, \dots, n$. 由此得到 n 个统计量 $X_{(1)}, X_{(2)}, \dots, X_{(n)}$, 且他们满足 $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, 称 $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ 为该样本的**顺序统计量或次序统计量**.

- $X_{(1)} = \min \{X_1, X_2, \dots, X_n\}$ 称为最小顺序统计量
- $X_{(n)} = \max \{X_1, X_2, \dots, X_n\}$ 称为最大顺序统计量

4. 样本极差

➤ 统计量

$$R = X_{(n)} - X_{(1)}$$

称为样本极差，其观测值为 $r = x_{(n)} - x_{(1)}$

5. 样本 p 分位数

➤ 对于 $0 < p < 1$ ，统计量

$$M_p = \begin{cases} X_{([np+1])}, & np \text{ 不是整数,} \\ \frac{1}{2}(X_{(np)} + X_{(np+1)}), & np \text{ 是整数} \end{cases}$$

称为样本 p 分位数，这里 $[a]$ 为不超过 a 的最大整数. 当 $p=1/2$ 时，样本 p 分位数称做**样本中位数**.

6.4 抽样分布

➤ 统计量是随机变量，所以有对应的概率分布，称之为统计量的抽样分布

1. Γ 分布

定义6.2: 如果随机变量 X 具有概率密度

$$f(x; \alpha, \lambda) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

其中 $\alpha > 0, \lambda > 0$ 为常数，则称 X 服从参数为 α, λ 的 Γ 分布，记为 $X \sim \Gamma(\alpha, \lambda)$ 。

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$$

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$$

► 性质1: 如果随机变量 $X \sim \Gamma(\alpha, \lambda)$, 那么 $E(X) = \frac{\alpha}{\lambda}$,

$$D(X) = \frac{\alpha}{\lambda^2}.$$

证: 由 Γ 函数的性质得

$$E(X) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} x^\alpha e^{-\lambda x} dx = \frac{\Gamma(\alpha+1)}{\lambda \Gamma(\alpha)} = \frac{\alpha}{\lambda}$$

又因为

$$E(X^2) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} x^{\alpha+1} e^{-\lambda x} dx = \frac{\Gamma(\alpha+2)}{\lambda^2 \Gamma(\alpha)} = \frac{\alpha(\alpha+1)}{\lambda^2}$$

所以

$$D(X) = E(X^2) - [E(X)]^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha}{\lambda^2}$$

► 性质2: 如果随机变量 $X_1 \sim \Gamma(\alpha_1, \lambda)$, $X_2 \sim \Gamma(\alpha_2, \lambda)$, 且

X_1 和 X_2 相互独立, 则 $X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$

证: 由两个独立随机变量和得卷积公式, $X_1 + X_2$ 的概率密度为

$$\begin{aligned} f(z) &= \int_{-\infty}^{+\infty} f_{X_1}(x; \alpha_1, \lambda) f_{X_2}(z-x; \alpha_2, \lambda) dx \\ &= \frac{\lambda^{\alpha_1}}{\Gamma(\alpha_1)} \int_0^{+\infty} x^{\alpha_1-1} e^{-\lambda x} f_{X_2}(z-x; \alpha_2, \lambda) dx \end{aligned}$$

其中

$$f_{X_2}(z-x; \alpha_2, \lambda) = \begin{cases} \frac{\lambda^{\alpha_2}}{\Gamma(\alpha_2)} (z-x)^{\alpha_2-1} e^{-\lambda(z-x)}, & x < z \\ 0, & x \geq z \end{cases}$$

所以, 当 $z \leq 0$ 时, $f(z)=0$; 当 $z > 0$ 时,

$$\begin{aligned}
 f(z) &= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z x^{\alpha_1-1} e^{-\lambda x} (z-x)^{\alpha_2-1} e^{-\lambda(z-x)} dx \\
 &= \frac{\lambda^{\alpha_1+\alpha_2} e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z x^{\alpha_1-1} (z-x)^{\alpha_2-1} dx \\
 &\stackrel{x=tz}{=} \frac{\lambda^{\alpha_1+\alpha_2} z^{\alpha_1+\alpha_2-1} e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 x^{\alpha_1-1} (1-t)^{\alpha_2-1} dx \\
 &= \frac{B(\alpha_1, \alpha_2) \lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z^{\alpha_1+\alpha_2-1} e^{-\lambda z} \\
 &= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)} z^{\alpha_1+\alpha_2-1} e^{-\lambda z}
 \end{aligned}$$

显然有, $X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$

设随机变量 $X_1, X_2, \dots, X_m$ 相互独立, 且 $X_i \sim \Gamma(\alpha_i, \lambda)$, 则

$$X_1 + X_2 + \dots + X_m \sim \Gamma(\alpha_1 + \alpha_2 + \dots + \alpha_m, \lambda)$$

2. χ^2 分布

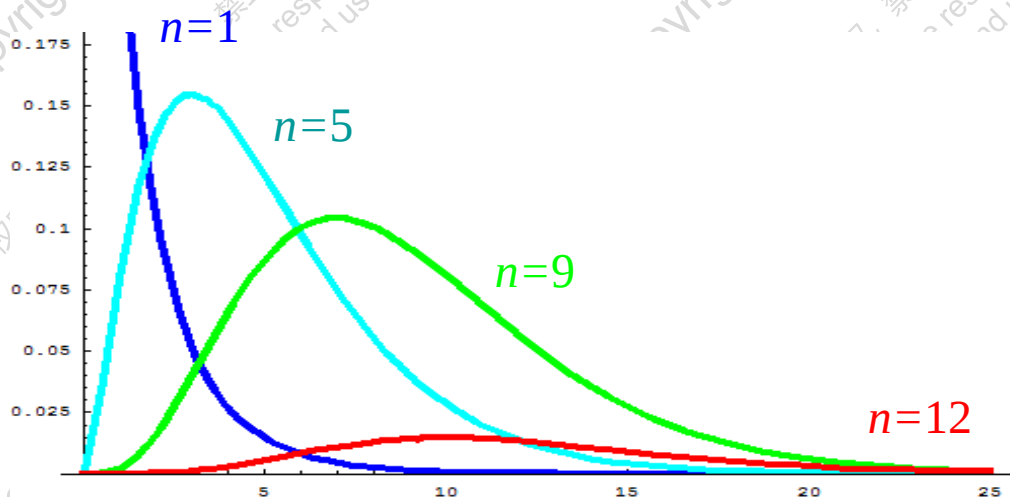
在 Γ 分布中取 $\alpha = \frac{n}{2}, \lambda = \frac{1}{2}$, Γ 分布就是自由度为 n 的卡方分布

定义6.3: 若随机变量 Z 具有概率密度

$$\chi^2(x; n) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0, \\ 0, & x \leq 0 \end{cases}$$

则称 Z 服从自由度为 n 的 χ^2 分布, 记作 $Z \sim \chi^2(n)$, 随机变量 Z 称为卡方变量.

$n = 1, 5, 9, 12$ 时, χ^2 分布的概率密度函数曲线.



当 $n = 2$ 时, 其密度函数是参数 $1/2$ 的指数分布

- 性质1: 如果随机变量 $Z \sim \chi^2(n)$, 那么 $E(Z) = n$,
 $D(Z) = 2n$.
- 性质2: 如果随机变量 $Z_1, Z_2, \dots, Z_m$ 相互独立, 且
 $Z_i \sim \chi^2(n_i) (i = 1, 2, \dots, m)$, 则 $\sum_{i=1}^m Z_i \sim \chi^2(n_1 + n_2 + \dots + n_m)$.

定理6.4 设随机变量 $X_1, X_2, \dots, X_n$, 且都服从标准正态分布 $N(0,1)$, 则随机变量 $\chi^2 = \sum_{i=1}^n X_i^2$ 服从自由度为 n 的卡方分布, 即 $\chi^2 \sim \chi^2(n)$.

定理6.4 设随机变量 $X_1, X_2, \dots, X_n$, 且都服从标准正态分布 $N(0,1)$, 则随机变量 $\chi^2 = \sum_{i=1}^n X_i^2$ 服从自由度为 n 的卡方分布, 即 $\chi^2 \sim \chi^2(n)$.

证: 若 $X \sim N(0,1)$, 则 X^2 的概率密度为

$$f(y) = \begin{cases} \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}}, & y > 0, \\ 0, & y \leq 0 \end{cases}$$

与

$$\chi^2(x; n) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0, \\ 0, & x \leq 0 \end{cases}$$

比较, 知 $X^2 \sim \chi^2(1)$, 因此, 结论成立.

例 4 设随机变量 $X \sim N(\mu, \sigma^2)$, $(X_1, \dots, X_n)$ 是来自总体 X 的样本, 求 $Y = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$ 的分布

解 因为 $X_i \sim N(\mu, \sigma^2) (i=1, 2, \dots, n)$, 且相互独立

所以 $\frac{X_i - \mu}{\sigma} \sim N(0, 1) (i=1, 2, \dots, n)$, 且相互独立

故

$$\begin{aligned} Y &= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \\ &= \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n) \end{aligned}$$

3. t 分布

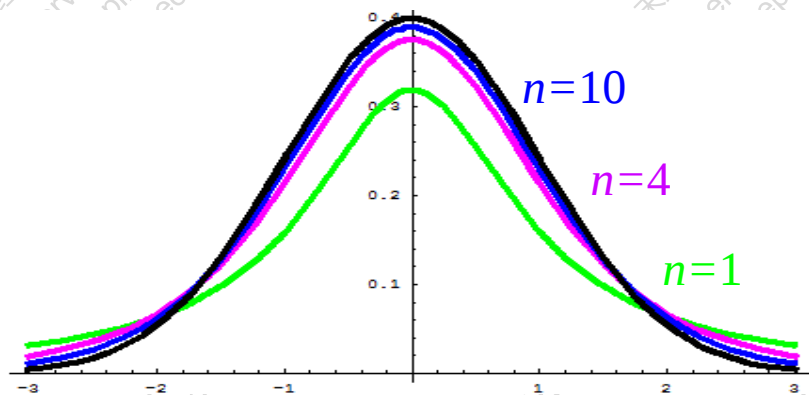
定义6.3: 若随机变量 T 具有概率密度函数

$$t(x; n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n} \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < x < +\infty$$

则称 T 服从**自由度为 n 的 t 分布**, 记作 $T \sim t(n)$.

t 分布又称为**学生氏(Student)分布**

$n = 1, 4, 10$ 时, t 分布的概率密度曲线, 和标准正态分布的概率密度曲线.



t 分布 $\xrightarrow{n \rightarrow +\infty}$

标准正态分布

定理6.5 若 $X \sim N(0, 1)$, $Y \sim \chi^2(n)$, 并且 X 与 Y 相互独立, 则

$$T = \frac{X}{\sqrt{Y/n}} \sim t(n)$$

证: 令 $Z = \sqrt{Y/n}$, 由 Y 的概率密度公式及因为2.2节中随机变量概率密度公式, 得 Z 得概率密度为

$$f_2(y) = \begin{cases} \frac{n^{\frac{n}{2}}}{2^{\frac{n}{2}-1} \Gamma\left(\frac{n}{2}\right)} y^{n-1} e^{-\frac{ny^2}{2}}, & y > 0, \\ 0, & y \leq 0 \end{cases}$$

于是, X 与 Z 相互独立, 由商 $T = X/Z$ 的概率密度公式有

$$g(t) = \int_{-\infty}^{+\infty} |y| f_1(ty) f_2(y) dy$$

$$t(x; n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n} \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$$

$$g(t) = \int_{-\infty}^{+\infty} |y| f_1(ty) f_2(y) dy$$

$$= \frac{n^{\frac{n}{2}}}{\sqrt{\pi} 2^{\frac{n-1}{2}} \Gamma\left(\frac{n}{2}\right)} \int_0^{+\infty} y^n e^{-\frac{(n+t^2)y^2}{2}} dy$$

$$= \frac{1}{\sqrt{n\pi} \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \int_0^{+\infty} s^{\frac{n-1}{2}} e^{-s} ds$$

$$= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad (-\infty \leq t \leq +\infty)$$

4. F 分布

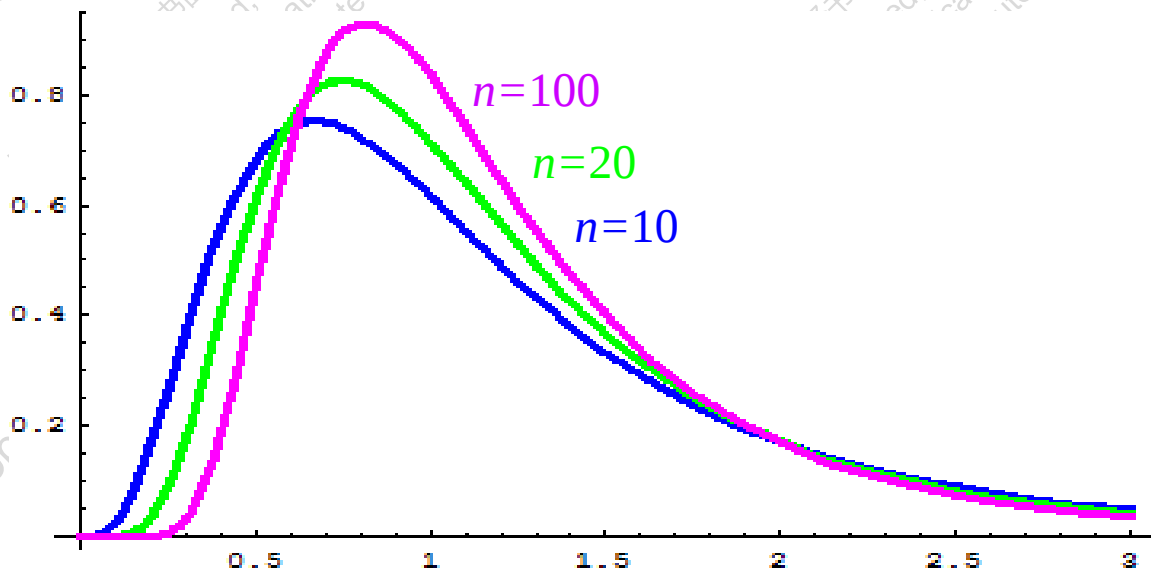
定义6.5: 若随机变量 F 具有概率密度服从

$$f(x; n, m) = \begin{cases} \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1} \left(1 + \frac{n}{m}x\right)^{-\frac{n+m}{2}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

则称 F 服从自由度为 (n, m) 的 F 分布,

记作 $F \sim F(n, m)$.

$(n, m) = (10, 10), (20, 10), (100, 10)$ 时, F 分布的概率密度曲线.



定理6.6 若 $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$, 且 X 与 Y 相互独立, 则

$$F = \frac{X/n}{Y/m} \sim F(n, m)$$

推论 若 $F \sim F(n, m)$, 则

$$\frac{1}{F} \sim F(m, n)$$

5.分位数

定义6.6: 设随机变量 X 的分布函数为 $F(x)$, 给定 $0 < p < 1$,
若有 x_p 满足 $P \{ X \leq x_p \} = F(x_p) = p$,
则称 x_p 为分布 $F(x)$ (或随机变量 X)的下侧 p 分位数

给定 $0 < \alpha < 1$, 若有 y_α 满足 $P \{ X > y_\alpha \} = 1 - F(y_\alpha) = \alpha$,

则称 y_α 为分布 $F(x)$ (或随机变量 X)的上侧 α 分位数

由定义可知, $y_\alpha = x_{1-\alpha}$; $x_p = y_{1-p}$.

记

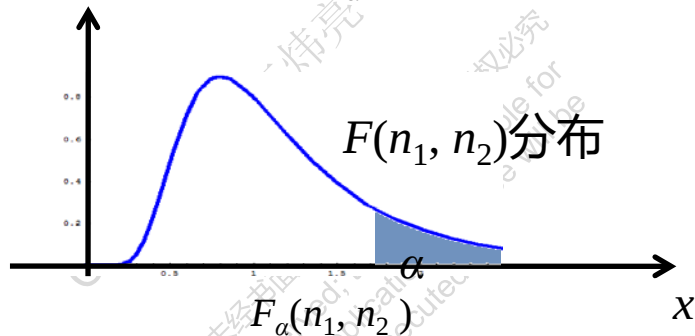
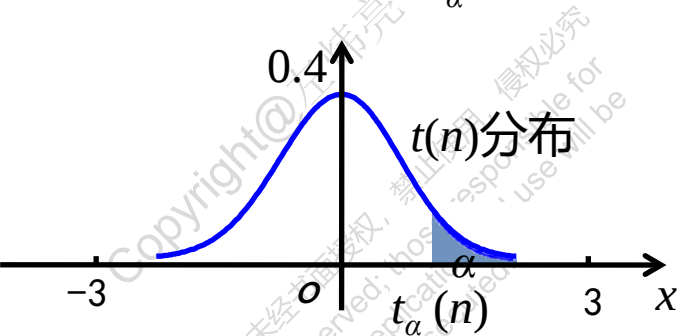
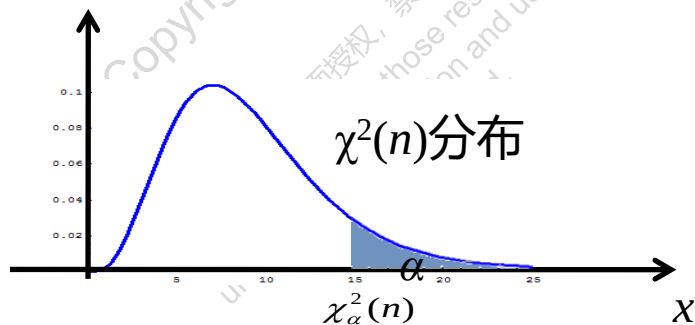
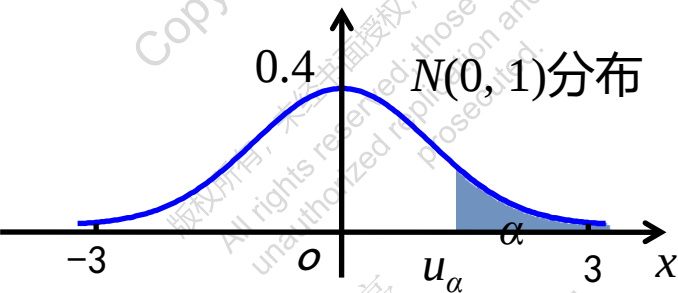
u_α —— $N(0, 1)$ 分布的上侧 α 分位数

$\chi_\alpha^2(n)$ —— $\chi^2(n)$ 分布的上侧 α 分位数

$t_\alpha(n)$ —— $t(n)$ 分布的上侧 α 分位数

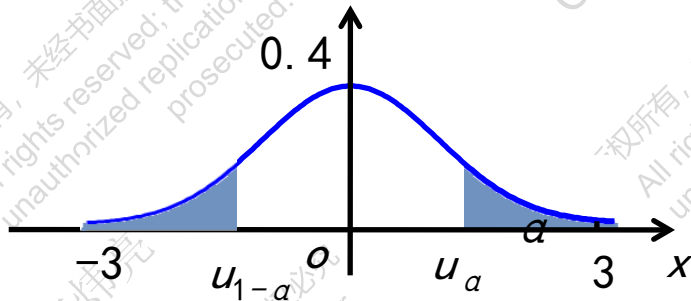
$F_\alpha(n_1, n_2)$ —— $F(n_1, n_2)$ 分布的上侧 α 分位数

上侧 α 分位数图示



注 1) 由 $N(0, 1)$ 分布, $t(n)$ 分布的对称性可:

$$u_{1-\alpha} = -u_{\alpha}, \quad t_{1-\alpha}(n) = -t_{\alpha}(n).$$



2) 当 n 充分大时, 因 $t(n) \approx N(0, 1)$, 故有

$$t_{\alpha}(n) \approx u_{\alpha}.$$

$$3) \quad F_{\alpha}(n, m) = \frac{1}{F_{1-\alpha}(m, n)}$$

证 3) 由定理6.6的推论知 $F \sim F(n, m)$, $\frac{1}{F} \sim F(m, n)$

由上侧 α 分位数定义有

$$\alpha = P\{F > F_{\alpha}(n, m)\} \quad \text{两边同时取倒数}$$

$$= P\left\{\frac{1}{F} < \frac{1}{F_{\alpha}(n, m)}\right\}$$

则

$$P\left\{\frac{1}{F} > \frac{1}{F_{\alpha}(n, m)}\right\} = 1 - \alpha$$

故, 按照定义有

$$\frac{1}{F_{\alpha}(n, m)} = F_{1-\alpha}(m, n)$$

P138页例题6.10-6.12

Copyright@左炜亮

版权所有，未经书面授权，禁止使用，侵权必究
All rights reserved; those responsible for
unauthorized replication and use will be
prosecuted.

Copyright@左炜亮

版权所有，未经书面授权，禁止使用，侵权必究
All rights reserved; those responsible for
unauthorized replication and use will be
prosecuted.

Copyright@左炜亮

版权所有，未经书面授权，禁止使用，侵权必究
All rights reserved; those responsible for
unauthorized replication and use will be
prosecuted.

Copyright@左炜亮

版权所有，未经书面授权，禁止使用，侵权必究
All rights reserved; those responsible for
unauthorized replication and use will be
prosecuted.

6.正态总体的抽样分布

定理6.7 设 $X_1, X_2, \dots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本, $\bar{X}$ 为样本均值, S^2 为样本方差, 则

$$1) \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$2) \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi^2(n-1)$$

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

3) $\bar{X}$ 与 S^2 相互独立.

定理6.8 设 $X_1, \dots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本, $\bar{X}$ 为样本均值, S^2 为样本方差, 则

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim T(n-1)$$

证: 由定理6.7知 $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)$, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$,

且 $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$ 与 $\frac{(n-1)S^2}{\sigma^2}$ 相互独立, 再由定理6.5知

$$T = \frac{\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{(n-1)}}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n-1)$$

定理6.9 设 $X_1, X_2, \dots, X_n$ 与 $Y_1, Y_2, \dots, Y_m$ 分别为来自正态总体 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ 的样本且两组样本相互独立, 记

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\bar{Y} = \frac{1}{m} \sum_{j=1}^m Y_j$$

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$S_2^2 = \frac{1}{m-1} \sum_{j=1}^m (Y_j - \bar{Y})^2$$

则

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}} \sim t(n+m-2)$$

证：由定理条件及定理6.7的结论（1）知

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \left(\frac{1}{n} + \frac{1}{m}\right)\sigma^2\right), \quad \text{即}$$

$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1),$$

再由定理条件、定理6.7的结论（2）及卡方分布的性质2知

$$V = \frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2} \sim \chi^2(n+m-2)$$

且由定理3.1及定理6.7的结论（3）知U与V相互独立，从而根据定理6.5知

$$T = \frac{U}{\sqrt{V/(n+m-2)}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}} \sim t(n+m-2)$$

注：定理中两个正太总体的方差相同

定理6.10 设 $X_1, X_2, \dots, X_n$ 与 $Y_1, Y_2, \dots, Y_m$ 分别为来自正态总体 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ 的样本, 且两组样本相互独立, S_1 和 S_2 如定理6.9定义, 则

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(n-1, m-1)$$

证: 由定理条件、定理6.7知 $\frac{(n-1)S_1^2}{\sigma_1^2}$ 和 $\frac{(m-1)S_2^2}{\sigma_2^2}$ 相互独立, 且

$$\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi^2(n-1), \quad \frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi^2(m-1)$$

从而由定理6.6知

$$F = \frac{\frac{(n-1)S_1^2}{\sigma_1^2} / (n-1)}{\frac{(m-1)S_2^2}{\sigma_2^2} / (m-1)} = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(n-1, m-1)$$

例5：在总体 $N(52, 6.3^2)$ 中，随机抽取一个容量为36的样本，求样本均值落在50.8到53.8之间的概率。

解：因为 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_n$ 服从正态分布

所以 $\bar{X} \sim N(52, 6.3^2 / 36)$

$$\begin{aligned} \text{故 } P\{50.8 \leq \bar{X} \leq 53.8\} &= \Phi\left(\frac{53.8 - 52}{6.3 / 6}\right) - \Phi\left(\frac{50.8 - 52}{6.3 / 6}\right) \\ &= \Phi(1.7143) - \Phi(-1.1429) \\ &= 0.8302 \end{aligned}$$

例6: 设总体 X 的概率密度函数为 $f(x) = \begin{cases} |x|, & |x| < 1; \\ 0, & |x| \geq 1. \end{cases}$

$(X_1, X_2, \dots, X_{50})$ 为来自总体 X 的样本, 求

(1) $\bar{X}$ 的数学期望和方差;

(2) $P\{|\bar{X}| > 0.02\}$.

解: (1) $E(\bar{X}) = E(X) = \int_{-1}^1 x|x|dx = 0$

$$D(\bar{X}) = \frac{1}{50} D(X) = \frac{1}{50} E(X^2) = \frac{1}{50} \int_{-1}^1 x^2|x|dx = \frac{1}{100}$$

(2) 因为 $\bar{X} \xrightarrow{d} N(0, 0.01)$ (根据中心极限定理)

$$\text{所以 } P\left\{\left|\bar{X}\right| > 0.02\right\} = 1 - P\left\{\left|\bar{X}\right| \leq 0.02\right\}$$

$$\approx 2\left(1 - \Phi\left(\frac{0.02 - 0}{0.1}\right)\right)$$

$$= 2(1 - \Phi(0.2))$$

$$= 0.8414$$

概率统计辅导书P107

Copyright@左炜亮
版权所有，未经书面授权，禁止使用，侵权必究
All rights reserved; those responsible for
unauthorized replication and use will be
prosecuted.

Copyright@左炜亮
版权所有，未经书面授权，禁止使用，侵权必究
All rights reserved; those responsible for
unauthorized replication and use will be
prosecuted.

Copyright@左炜亮
版权所有，未经书面授权，禁止使用，侵权必究
All rights reserved; those responsible for
unauthorized replication and use will be
prosecuted.

Copyright@左炜亮
版权所有，未经书面授权，禁止使用，侵权必究
All rights reserved; those responsible for
unauthorized replication and use will be
prosecuted.